公告本

申請日期	90.7.17.
案號	90117207
類別	G10L 15/06

（以上各欄由本局填註）

A4
C4

514867

發明 ~~新型~~ 專利說明書

一、發明 ~~新型~~ 名稱	中文	建構一與演說者無關之語音辨識系統之語音樣板之方法及裝置
	英文	"METHOD AND APPARATUS FOR CONSTRUCTING VOICE TEMPLATES FOR A SPEAKER-INDEPENDENT VOICE RECOGNITION SYSTEM"
二、發明 ~~創作~~ 人	姓名	畢寧 NING BI
	國籍	中國
	住、居所	美國加州聖地牙哥市和風道廣場14209號
三、申請人	姓名（名稱）	美商奎康公司 QUALCOMM INCORPORATED
	國籍	美國
	住、居所（事務所）	美國加州聖地牙哥市摩豪斯大道5775號
	代表人姓名	菲力普 R. 華德渥斯 PHILIP R. WADSWORTH

（由本局填寫）

承辦人代碼：	A6 B6
大　　類：	
ＩＰＣ分類：	

本案已向：

國（地區）	申請專利，申請日期：	案號：	，□有　□無主張優先權
美國	2000 年 07 月 13 日	09/615,572	☑有　□無主張優先權

有關微生物已寄存於：　　，寄存日期：　　，寄存號碼：

（請先閱讀背面之注意事項再填寫本頁各欄）

裝　訂　線

五、發明說明（ 1 ）

發明背景

I. 發明範疇：

本發明一般而言係關於通訊範疇，更特定而言，係關於與演說者無關的語音辨識系統的語音樣板。

II. 背景：

語音辨識(VR)代表賦與機器模擬的智慧來辨識使用者或使用者語音指令，並便於與該機器的人機介面之最重要的技術之一。VR也代表人類說話瞭解的關鍵技術。使用技術來由一聲音說話信號還原一語言訊息的系統即稱之爲語音辨識器。此處所使用的該名詞"語音辨識器"通常代表任何說話使用者介面致能的裝置。一語音辨識器通常包含一聲音處理器及一字元解碼器。該聲音處理器擷取一系列的包含資訊的特徵或向量，其必須來達到進入的原始演說的VR。該字元解碼器解碼該特徵或向量的序列來產生一有意義及所需要的輸出格式，例如對應於該輸入發音的一系列語言字元。

該聲音處理器代表一語音辨識器中的前端演說分析子系統。回應於一輸入演說信號，該聲音處理器提供一適當的代表來特徵化該隨時間變化的演說信號。該聲音處理器必須捨棄不相關的資訊，例如背景雜訊，通道失眞，擴音器特性，及演說方式等。有效率的聲音處理可使得語音辨識器具有加強的聲音分辨能力。目前爲止，要分析的一有用特徵爲該短時間頻譜外形。兩種常用的頻譜分析技術來特徵化該短時間頻譜外形爲線性預測編碼(LPC)及濾波器庫

五、發明說明（ 2 ）

爲主的頻譜模型化。範例性 LPC 技術係描述於美國專利編號 5,414,796 中，其授權給本發明的受讓人，在此完全引用做爲參考，以及 L.B. Rabiner 及 R.W. Schafer 所著的 Digital Processing of Speech Signals 396-453 (1978)，其也在此完全引用做爲參考。

VR 的使用(通常也稱之爲 speech recognition)，爲了安全的理由而日漸重要。舉例而言， VR 可用來取代無線電話鍵盤的人爲按鈕動作。此係特別地有用，當使用者在開車時要撥打電話時。當使用沒有 VR 的電話時，該駕駛者必須由駕駛盤移出一隻手，並注視著鍵盤來按鈕及撥打電話。這些動作增加了車禍的可能性。一種語音啓動電話(及設計有語音辨識的電話)，其將允許該駕駛者來撥打電話，而仍可注視著前方道路。一種車用免持聽筒系統可額外地允許駕駛者來在打電話時成維持雙手在方向盤上。

語音辨識裝置可分類成演說者相關(SD)或演說者無關(SI)之裝置。擴音器相關裝置較爲常見，其被訓練爲辨識來自特殊使用者的指令。相反地，擴音器無關裝置能夠接受來自任何使用者的聲音指令。爲了增加一給定的 VR 系統之效能，不論是演說者相關或演說者無關，皆需要訓練來建立該系統有效的參數。換言之，該系統需要在其最佳化運作之前來學習。

一演說者相關 VR 裝置基本上以兩個階段運作，即訓練階段及辨識階段。在訓練階段中，該 VR 系統提示使用者說出系統詞彙中的字元一次或兩次(基本上爲兩次)，所以

五、發明說明（ 3 ）

該系統可學習到使用者對這些特殊字元或片語的語音特性。對於一車用免持聽筒的範例詞彙可包含鍵盤上的數字；關鍵字"打電話"，"傳送"，"撥號"，"取消"，"清除"，"加入"，"刪除"，"歷史"，"程式"，"是"及"否"；及經常撥打的同事，朋友或親人之預先定義號碼的名字。一旦完成訓練，該使用者可說出該訓練過的字元來在辨識階段啓始打電話，而 VR 裝置藉由比較說出的發音與先前訓練的發音(儲存爲樣板)，並採用最佳的匹配來辨識。舉例而言，如果名字"John"爲訓練的名字之一，該使用者可藉由說出片語"Call John"來啓始打電話給 John。該 VR 系統可辨識字元"Call"及"John"，並可撥出該使用者先前輸入的 John 的電話號碼。此即爲訓練的系統及方法。

一演說者無關的 VR 裝置也使用一訓練樣板，其包含一預先定義尺寸的預錄詞彙(如某些控制字元，由 0 到 9 的數字，是及否)。大量的演說者(如 100)必須錄製說出詞彙中的每個字元。

習用上，演說者無關的 VR 樣板係由比較包含由第一組演說者(基本上 100 名演說者)所說出的字元之測試資料庫與包含由第二組演說者(與第一組同樣多)說出相同字元的一訓練資料庫。一字元，由一使用者說出，其基本上稱之爲一發音。每個該訓練資料庫的發音係先經過時間正規化，然後在測試與該測試資料庫的發音之收斂性之前被量化(基本上係根據已知的技術來向量量化)。但是，該時間正規化技術依賴僅由個別訊框(一發音的週期性段落)得到

五、發明說明（ 4 ）

的資訊，其具有與先前訊框的最大差異。其較佳地是來提供一方法來使用一給定發音中更多的資訊來構建演說者無關的 VR 樣板之方法。其進一步需要來增加根據發音形式來構建演說者無關的 VR 樣板之習用技術的準確性或收斂性。因此，其有需要一種方法來建構演說者無關的語音辨識樣板，其提供加強的準確性，並使用該發音中的較大量的資訊。

發明概要

本發明係關於一種建構演說者無關的語音辨識樣板之方法，其提供加強的準確性及使用該發音中較大量的資訊。因此，在本發明的一方面中，其提供一種產生說話樣板的方法，用於一演說者無關的語音辨識系統。該方法較佳地是包含分段每個一第一複數個發音之發音來對每個發音產生複數個時間叢集的段落，每個時間叢集化的段落係由一頻譜平均值來代表；量化所有該第一複數個發音的該複數個頻譜平均來產生複數個樣板向量；比較該複數個樣板向量與一第二複數個發音的每一個來產生至少一個比較結果；匹配該第一複數個發音與該複數個樣板向量，如果該至少一個比較結果超過至少一個預定的臨限值，以產生一最佳化匹配路徑結果；根據該最佳化匹配路徑結果在時間中區隔化該第一複數個發音；並重複該量化，比較，匹配及區隔化，直到該至少一個比較結果並未超過任何至少一個預定的臨限值。

圖式簡單說明

五、發明說明（ 5 ）

圖 1 所示爲用於建構及實施演說者無關的語音辨識之語音樣板的系統方塊圖。

圖 2 所示爲可用於圖 1 的系統中之語音辨識系統的方塊圖。

圖 3 所示爲由一語音辨識子系統，例如圖 2 的子系統，所執行方法的流程圖，藉以辨識輸入語音樣本。

圖 4 所示爲可用於圖 1 的系統之樣板建構子系統的方塊圖。

圖 5 所示爲可用於圖 1 的系統之樣板建構子系統的方塊圖。

圖 6 所示爲由一樣板建構子系統，例如圖 4 的子系統或圖 5 的子系統，所執行的方法步驟之流程圖，用以建構語音樣板。

較佳具體實施例之詳細說明

根據一具體實施例，如圖 1 所示，一用以建構及實施演說者無關的語音辨識之語音樣板的系統 10，其包含一演說者無關樣板建構子系統 12 及一語音辨識子系統 14。該演說者無關樣板建構子系統 12 係結合於該語音辨識子系統 14。

演說者無關語音樣板係由該演說者無關樣板建構子系統 12 所建構，如下述圖 4-6。該樣板係提供給該語音辨識子系統 14 來用於辨識來自一使用者的輸入語音，如下述之圖 2-3。

根據一具體實施例，如圖 2 所示，一語音辨識子系統

五、發明說明（ 6 ）

100 包含一類比到數位轉換器(A/D) 102 ，一前端聲音處理器 104，一特徵擷取器 106，一語音樣板資料庫 108，圖樣比較邏輯 110，及決策邏輯 112。在一特殊具體實施例中，該聲音處理器 104 及該特徵擷取器 106 係實施成一個裝置，例如一參數擷取器。在一具體實施例中，該聲音處理器 104 包含一頻率分析模組 114。在一具體實施例中，該特徵擷取器 106 包含一終點偵測器 116，一時間叢集語音分段模組 118，及一語音位準正規化器 120。

該 A/D 102 係耦合於該聲音處理器 104。該聲音處理器 104 係耦合於該特徵擷取器 106。在一具體實施例中，該特徵擷取器 106 內，該終點偵測器 116 係耦合於該時間叢集語音分段模組 118，其耦合於該振幅量化器 120。該特徵擷取器 106 耦合於該圖樣比較邏輯 110。該圖樣比較邏輯 110 耦合於該樣板資料庫 108 及該決策邏輯 112。

該語音辨識子系統 100 可存在於像是無線電話或一車用免持聽筒。一使用者(未示出)說出一字元或片語，即產生一語音信號。該語音信號即利用一習用的換能器(未示出)來轉換到一電子語音信號 s(t)。該語音信號 s(t) 即提供給 A/D 102 ，其根據一已知的取樣方法，例如像是脈衝編碼調變(PCM)， A-法則或 μ-法則來轉換該語音信號到數位化的語音樣本 s(n)。

該語音樣本 s(n) 係提供給該聲音處理器 104 來決定參數。該聲音處理器 104 產生一組參數，其模型化該輸入語音信號 s(t) 的特徵。該參數可根據任何已知的一些語音參

五、發明說明（ 7 ）

數決定技術來決定，其包含像是語音編碼器編碼，離散傅立葉轉換(DFT)爲主的頻譜係數(如快速傅立葉轉換(FFT)爲主的頻譜係數)，線性預測係數(LPCs)，或巴克等級分析，如前述美國專利編號 5,414,796 ，及 Lawrence Rabiner 及 Biing-Hwang Juang 所著 Fundamentals of Speech Recognition (1993)中所述。該組參數較佳地是爲以訊框爲主(分段成週期性訊框)。該聲音處理器 104 可實施爲一數位信號處理器(DSP)。該 DSP 可包含一語音編碼器。另外，該聲音處理器 104 可實施爲一語音編碼器。

每個參數訊框即提供給該特徵擷取器 106 。在該特徵擷取器 106 中，該終點偵測器 116 使用該擷取的參數來偵測一發音(即一字元)的終點。在一具體實施例中，該終點偵測較佳地是根據美國專利申請編號 09/246,414 中所述的原理來進行，其於 1999 年 2 月 8 日立案，名爲"在有雜訊之下準確地找出語音終點的方法及裝置"("METHOD AND APPARATUS FOR ACCURATE ENDPOINTING OF SPEECH IN THE PRESENCE OF NOISE")，其授權給本發明的受讓人，在此完全引用做爲參考。根據此技術，該發音係相較於一第一臨限值，例如像是一信號對雜訊比(SNR)臨限值，以決定該發音的一第一啓始點及一第一終止點。在該第一啓始點之前的該發音的一部份即與一第二 SNR 臨限值比較，以決定該發音的一第二啓始點。在該第一終止點之後之該發音係與第二 SNR 臨界值比較，以決定該發音之第二終止點。該第一及第二 SNR 臨限值較佳地是週期性地重新計

五、發明說明（ 8 ）

算，而該第一 SNR 臨限值較佳地是超過該第二 SNR 臨限值。

該偵測的發音之頻率領域參數的訊框係提供給該時間叢集語音分段模組 118，其根據一具體實施例而實施一壓縮技術，其述於美國專利申請編號 09/225,891，於 1999 年 1 月 4 日提案，名爲"語音信號的分段及辨識之系統及方法"("SYSTEM AND METHOD FOR SEGMENTATION AND RECOGNITION OF SPEECH SIGNALS")，其授權給本發明的受讓人，在此完全引用做爲參考。根據此技術，該頻率領域參數中的每個語音訊框係以關於該語音訊框的至少一個頻譜值來代表。然後一頻譜差異值即對每對相鄰的訊框來決定。該頻譜差異值代表關於該配對中兩個訊框的頻譜值之間的差異。一初始叢集邊界係設定在每對相鄰訊框之間，產生參數中的叢集，及將一變化值指定給每個叢集。該變化值較佳地是等於該決定的頻譜差異值之一。然後計算複數個叢集結合參數，每個該叢集結合參數係關於一對相鄰的叢集。一最小的叢集結合參數係選自複數個叢集結合參數。然後一結合的叢集即由取消關於該最小叢集結合參數的該叢集之間的叢集邊界來形成。該結合的變化值代表指定給關於該最小叢集結合參數的該叢集之變化值。該處理較佳地是重複進行，藉以形成複數個結合的叢集，且該分段的語音信號可較佳地是根據該複數個結合的叢集來形成。

本技藝的專業人士將可瞭解，該時間叢集語音分段模組

五、發明說明（ 9 ）

118 可由其它裝置來取代，例如像是一時間正規化模組。但是，專業人士亦可瞭解到，因爲該時間叢集語音分段模組 118 結合具有最小差異的訊框成爲叢集，其係相較於先前的訊框，並使用平均值而非個別的訊框，該時間叢集語音分段模組 118 使用該處理的發音中更多的資訊。其亦可瞭解到，該時間叢集語音分段模組 118 較佳地是配合圖樣比較邏輯 110 來使用，其利用本技藝中所熟知的動態時間扭曲(DTW)，如下所述。

該叢集平均係提供給該語音位準正規化器 120。在一具體實施例中，該語音位準正規化器 120 藉由指定給每個叢集平均每個通道兩個位元來量化該語音振幅(即每個頻率兩個位元)。在另一具體實施例中，其中頻譜係數被擷取，該語音位準正規化器 120 並未用於量化該叢集平均，其爲專業人士所能瞭解。由該語音位準正規化器 120 所產生的輸出係由該特徵擷取器 106 提供給該圖樣比較邏輯 110。

該語音辨識子系統 100 的所有詞彙字元的一組樣板係永久地儲存在該樣板資料庫 108。該組樣板較佳地是爲一組演說者無關的樣板，其以如下所述的一演說者無關的樣板建構子系統所建構。該樣板資料庫 108 較佳地是由任何習用的非揮發性媒體形式來實施，例如像是快閃記憶體。此允許該樣板來保持在該樣板資料庫 108 中，當該語音辨識子系統 100 的電源被關掉時。

該圖樣比較邏輯 110 比較來自該特徵擷取器 106 的向量

五、發明說明（ 10 ）

與儲存在樣板資料庫 108 中的所有樣板。在該向量與所有儲存在樣板資料庫 108 中的樣板之間的比較結果或距離將提供給該決策邏輯 112 。該決策邏輯 112 自該樣板資料庫 108 中選出最爲匹配該向量的樣板。另外，該決策邏輯 112 可使用一習用的"N-最佳"選擇演算法，其在一預定的匹配臨限值之內選出 N 個最接近的匹配。然後該使用者即被詢問是要那一個選擇。該決策邏輯 112 的輸出爲在該詞彙中決定出所說的字元。

在一具體實施例中，該圖樣比較邏輯 110 及該決策邏輯 112 使用一 DTW 技術來測試收斂性。該 DTW 技術在本技藝中爲人所熟知，並述於 Lawrence Rabiner 及 Biing-Hwang Juang 所著的 Fundamentals of Speech Recognition 200-238 (1993)。其在此完全引用做爲參考。根據該 DTW 技術，一格架由繪出要測試的該發音的一時間序列，對於儲存在該樣板資料庫 108 中的每個發音之時間序列所形成。然後正在測試的發音即與該樣板資料庫 108 中的每個發音進行比較，點對點方式(例如每 10 ms)，一次一個發音。對於樣板資料庫 108 中的每個發音，正在測試的發音被調整或"扭曲"在時間中，其係在特殊的點被壓縮或擴張，直到達到與該樣板資料庫 108 中的發音最可能接近的匹配。在時間中的每個點上，該兩個發音進行比較，其爲在該點(零成本)宣告匹配，或宣告一不匹配。在一特殊點上的一不匹配的事件中，正在測試的發音即被壓縮，擴張，或如果需要的話被錯配。該處理一直持續到該兩個發音已經彼此

五、發明說明（ 11 ）

完全地比較過。其有可能爲大量(典型會有數千)的不同調整的發音。該調整的發音具有最低成本的函數(即需要最少量的壓縮及/或擴張及/或錯配)被選出。以類似於 Viterbi 解碼演算法的方式，該選擇較佳地是由重新回顧該樣板資料庫 108 中的發音之每個點來進行，藉以決定具有最低整體成本的路徑。此允許該最低成本(即最爲匹配)的調整過發音，可以不需要重新排序產生每個不同調整的發音之所有可能之"強制"方法而決定出來。然後對於該樣板資料庫 108 中所有發音之該最低成本調整的發音即被比較，其中具有最低成本者即被選爲該儲存的發音中最爲接近匹配於該測試的發音。

該圖樣比較邏輯 110 及該決策邏輯 112 可較佳地實施爲一微處理器。該語音辨識子系統 100 可爲像是一 ASIC。該語音辨識子系統 100 的辨識準確性爲該語音辨識子系統 100 如何正確地辨識出該詞彙中說出的字元或片語之度量。舉例而言，一 95%辨識準確性代表該語音辨識子系統 100 可正確地在 100 次中辨識出該詞彙中的字元以 95 次。

根據一具體實施例，一語音辨識子系統(未示出)執行圖 3 所示的流程圖中的演算法步驟，以辨識語音輸入到該語音辨識子系統。在步驟 200 中，其提供輸入語音到該語音辨識子系統。然後控制流進行到步驟 202。在步驟 202 中，即偵測出一發音的終點。在一特殊具體實施例中，該發音的該終點係根據上述的美國專利申請編號 09/246,414 中所提充的技術來偵測，如上述配合圖 2 的控制流，即進

五、發明說明（ 12 ）

行到步驟 204。

在步驟 204 中，在該擷取的發音中執行時間叢集語音分段。在一特殊具體實施例中，所使用的該時間叢集語音分段技術即爲先前美國專利申請編號 09/225,891 中所述的技術，如圖 2 中所述。然後控制流進行到步驟 208。在步驟 206 中，係提供演說者無關的樣板來匹配於在步驟 204 中所產生的該語音叢集平均。該演說者無關樣板較佳地是根據下述參考圖 4-6 的技術所建構。然後控制流進行到步驟 208。在步驟 208 中，在一特殊發音的叢集與所有該演說者無關的樣板之間進行一 DTW 匹配，且該最接近的匹配樣板被選爲該辨識的發音。在一特殊具體實施例中，該 DTW 匹配係根據 Lawrence Rabiner 及 Biing-Hwang Juang 所著 Fundamentals of Speech Recognition 200-238 (1993) 中所述的技術來進行，其如圖 2 所示。本技藝的專業人士可以瞭解到，除了時間叢集語音分段之外的方法亦可在步驟 204 中執行。這種方法包含像是時間正規化。

根據一具體實施例，如圖 4 所示，一演說者無關的樣板建構子系統 300 包含一處理器 302 及一儲存媒體 304。該處理器 100 較佳地是爲一微處理器，但可爲任何習用種類的處理器，專屬處理器，數位信號處理器(DSP)，控制器或狀態機器。該處理器 302 耦合於該儲存媒體 304，其較佳地是實施爲快閃記憶體，EEPROM 記憶體，RAM 記憶體，設定用來保持韌體指令的 ROM 記憶體，執行在該處理器 302 上的一軟體模組，或任何其它習用種類的記憶體。該

五、發明說明（ 13 ）

演說者無關的樣板建構子系統 300 較佳地是實施爲執行 UNIX® 作業系統電腦。在另一具體實施例中，該儲存媒體 304 可爲板上的 RAM 記憶體，或該處理器 302 及該儲存媒體 304 可存在於一 ASIC。在一具體實施例中，該處理器 302 係用來執行包含於該儲存媒體 304 的一組指令，用以執行演算法步驟，例如下述參考圖 6 的步驟。

根據另一具體實施例，如圖 5 所示，一演說者無關的樣板建構子系統 400 包含一終點偵測器 402，時間叢集語音分段邏輯 404，一向量量化器 406，一收斂測試器 408，及 K 值平均語音分段邏輯 410。一控制處理器(未示出)可較佳地是用來控制該演說者無關的樣板建構子系統 400 所執行的遞迴數目。

該終點偵測器 402 係耦合於該時間叢集語音分段邏輯 404。該時間叢集語音分段邏輯 404 係耦合於該向量量化器 406。該向量量化器 406 係耦合於該收斂測試器 408，及該 K 值平均語音分段邏輯 410。該控制處理器較佳地是經由一控制匯流排(未示出)而耦合於該終點偵測器 402，該時間叢集語音分段邏輯 404，該向量量化器 406，該收斂測試器 408，及該 K 值平均語音分段邏輯 410。

要被訓練的一發音訓練樣本 $S_x(n)$ 係提供在訊框中給該終點偵測器 402。該訓練樣本較佳地是由一訓練資料庫(未示出)所提供，其中要訓練的發音被儲存。終點偵測器 402 偵測發音之啓始點及終止點。在一具體實施例中，該訓練資料庫包含 100 個字元，每個皆由 100 個不同的演說

五、發明說明（ 14 ）

者說出，而得到總共 10,000 個儲存的發音。在一具體實施例中，該終點偵測器 402 係根據前述美國專利申請編號 09/246,414 中所述的原理來運作，並參考上述圖 2。

該終點偵測器 402 提供該偵測的發音到該時間叢集語音分段邏輯 404。該時間叢集語音分段邏輯 404 在該偵測的發音上執行一壓縮演算法。在一具體實施例中，該時間叢集語音分段邏輯 404 係根據前述美國專利申請編號 09/225,891 中所述的原理來運作，並參考上述圖 2。在一具體實施例中，該時間叢集語音分段邏輯 404 壓縮該偵測的發音成二十個段落，每個段落包含一叢集平均。

該時間叢集語音分段邏輯 404 提供一給定字元的所有訓練發音的該叢集平均到該向量量化器 406。該向量量化器 406 量化該發音的該叢集平均(即對於相同字元的所有演說者)，並提供所得到的向量做爲該發音的可能演說者無關(SI)樣板給該收斂性測試器 408。該向量量化器 406 較佳地是根據任何許多已知的向量量化(VQ)技術來運作。不同的 VQ 技術係說明在像是 A. Gersho 及 R.M. Gray 所著 Quantization and Signal Compression (1992)。在一特殊具體實施例中，該向量量化器 406 產生四叢集向量。因此，例如每個段落係序列化提供給該向量量化器 406，其代表每個段落成爲四個叢集。每個叢集代表該特殊字元的每個演說者，且每個字元有數個叢集。根據一具體實施例，每個樣板有 8 個向量(4 個叢集乘以 20 個段落)。

該收斂測試器 408 比較該潛在的 SI 樣板與要測試的發音

五、發明說明（ 15 ）

之測試樣本 $S_y(n)$。該測試樣本係以訊框提供給該收斂測試器 408。該測試樣本較佳地是由一測試資料庫(未示出)來提供，其中儲存有要測試的發音。在一具體實施例中，該測試資料庫包含 100 個字元，每個係由 100 個不同的演說者來說出，成爲總共 10,000 個儲存的發音。該字元較佳地是與包含在訓練資料庫中爲相同的字元，但由 100 個不同的演說者說出。該收斂測試器 408 比較要訓練的發音之潛在的 SI 樣板，與要測試的發音之樣本。在一具體實施例中，該收斂測試器 408 係用來使用一 DTW 演算法來測試收斂性。該使用的 DTW 演算法較佳地是爲描述在 Lawrence Rabiner 及 Biing-Hwang Juang 所著的 Fundamentals of Speech Recognition 200-238 (1993) 及上述的圖 2 中。

在一具體實施例中，該收斂測試器 408 係用來分析資料庫中所有字元的結果之準確性及具有潛在的 SI 樣板的資料庫之變化。該變化先被檢查，而如果該變化低於一預定的臨限值時，該準確性即被檢查。該變化較佳地是對每個段落來計算，然後加總來產生一整體的變化值。在一特殊具體實施例中，該變化係由計算該 4 個叢集的最佳匹配的均方根誤差來得到。該均方根誤差技術爲本技藝中所熟知。該收斂性測試被定義爲準確，如果來自該測試資料庫的發音匹配於由該訓練資料庫所產生的潛在 SI 樣板(即如果辨識對於該資料庫中所有字元爲正確的話)。

該潛在的 SI 樣板也可由該向量量化器 406 提供給該 K 值平均語音分段邏輯 410。該 K 值平均語音分段邏輯 410 也

五、發明說明（ 16 ）

接收該訓練樣本，其較佳地是區隔成訊框。在該收斂性測試器 408 已執行第一收斂性測試之後，該變化或準確性的結果可低於該變化及準確性的預定臨限值。在一具體實施例中，如果該變化或準確性的結果低於該變化及準確性的預定臨限值時，即執行另一個遞迴。因此，該控制處理器指示該 K 值平均語音分段邏輯 410 來對該訓練樣本執行 K 值平均分段化，藉以產生如下述的分段語音訊框。根據該 K 值平均語音分段化，該訓練樣本係匹配於該潛在的 SI 樣本，較佳地是以一 DTW 技術，藉此產生如上述圖 2 之最佳路徑。然後該訓練樣本即根據該最佳路徑被分段化。舉例而言，該訓練樣本的前 5 個訊框可匹配於該可能 SI 樣板的第一訊框，接下來的 3 個訓練樣本的訊框可匹配於該可能 SI 樣板的第二訊框，而該訓練樣本的下 10 個訊框可匹配於該潛在的 SI 樣板的第三訊框。在此例中，該訓練樣本的前 5 個訊框將被分段成一個訊框，下三個訊框則被分段成一第二訊框，而下 10 個訊框將被分段成一第三訊框。在一具體實施例中，該 K 值平均語音分段邏輯 410 根據一範例性 K 值平均分段技術來執行該 K 值平均分段化，該技術描述於 Lawrence Rabiner 及 Biing-Hwang Juang 所著 Fundamentals of Speech Recognition 382-384 (1993)，其在此完全引用做爲參考。然後該 K 值平均語音分段邏輯 410 即提供該更新的叢集平均的訊框到該向量量化器 406，該向量量化器量化該叢集平均，並提供該結果向量(其包含新的潛在 SI 樣板)到該收斂性測試器 408，以執行另一個收斂

五、發明說明（ 17 ）

性測試。本技藝的一專業人士將可瞭解到此遞迴處理可依需要持續到達成在該預定臨限值之上的變化及準確性結果。

一旦該收斂性測試通過，該潛在的(現在爲最終的) SI樣板可較佳地用於一語音辨識子系統，例如圖 2 的語音辨識子系統。該最終的SI樣板將可儲存在圖 2 的樣板資料庫108中，或用於圖 3 的流程圖中的步驟 206。

在一具體實施例中，一演說者無關樣板建構子系統(未示出)執行示於圖 6 的流程圖中的方法步驟，以建構一發音的演說者無關的樣板。在步驟 500 中，一發音的訓練樣本係由一訓練資料庫獲得(未示出)。該訓練資料庫較佳地是包含大量的字元(如 100 個字元)，每個皆由大量的演說者說出(如每個字元 100 個演說者)。然後控制流進行到步驟 502。

在步驟 502 中，對於訓練樣本執行終點偵測，以偵測一發音。在一具體實施例中，該終點偵測係根據前述的美國專利申請編號 09/246,414 所述的技術執行，並參考上圖 2。然後控制流進行到步驟 504。

在步驟 504 中，時間叢集語音分段係對偵測的發音來執行，藉此壓縮該發音成數個段落，每個段落由一平均來代表。在一特殊具體實施例中，該發音係壓縮成 20 個段落，每個段落包含一叢集平均。在一具體實施例中，該時間叢集語音分段係根據前述的美國專利申請編號 09/225,891 所述的技術執行，並參考上圖 2。然後控制流

五、發明說明（ 18 ）

進行到步驟 506。

在步驟 506 中，該相同字元的所有演說者的訓練樣本的叢集平均皆爲向量量化。在特殊具體實施例中，該叢集平均係根據描述於 A. Gersho 及 R.M. Gray 所著 Vector Quantization and Signal Compression (1992) 中的許多已知技術中的一個來向量量化。在一特殊具體實施例中，產生 4 叢集向量。因此，例如每個段落係由 4 個叢集來代表。每個叢集代表該特殊字元的每個演說者，且每個字元有多個叢集。根據一具體實施例，每個樣板產生 80 個向量(4 個叢集乘以 20 個段落)。然後控制流進行到步驟 510。

在步驟 508 中，測試樣本由一測試資料庫(未示出)得到，以測試收斂性。該測試資料庫較佳地是包含在該訓練資料庫中含有的相同字元，其每個皆由大量的演說者所說出(例如每個發音有 100 個演說者)。然後控制流進行到步驟 510。

在步驟 510 中，該量化的向量係以潛在的 SI 樣板與該測試樣本來比較，以測試收斂性。在一具體實施例中，該收斂性測試爲一 DTW 演算法。該使用的 DTW 演算法可較佳地爲 Lawrence Rabiner 及 Biing-Hwang Juang 所著的 Fundamentals of Speech Recognition 200-238 (1993) 所述，並參考上圖 2。

在一具體實施例中，步驟 510 的收斂性測試分析所有資料庫中字位的結果之準確性及具有潛在 SI 樣板的資料庫之變化。該變化先被檢查，且如果該變化低於一預定的臨限

五、發明說明（ 19 ）

值時，該準確性即被檢查。該變化較佳地是以每個段落來計算，然後被加總來產生一整體的變化值。在一特殊具體實施例中，該變化係由計算該 4 個叢集的最佳匹配的均方根來取得。該均方根技術在本技藝中所熟知。該收斂性測試在由該測試資料庫產生的該可能 SI 樣板匹配於來自該訓練資料庫的該發音時，被定義爲準確(即如果辨識對於所有該資料庫中的字元爲正確)。然後控制流進行到步驟 512。

在步驟 512 中，如果步驟 510 的收斂性測試對於變化或準確性的結果低於變化及準確性的預定臨限值時，即執行另一個遞迴。因此，K 值平均語音分段即對該訓練樣本來執行。該 K 值平均語音分段匹配該訓練樣本於該潛在的 SI 樣板，其較佳地是以 DTW 技術執行，藉此產生一最佳化路徑，如參考上述之圖 2。然後該訓練樣本根據該最佳化路徑來分段。在一具體實施例中，該 K 值平均語音分段係根據 Lawrence Rabiner 及 Biing-Hwang Juang 所著 Fundamentals of Speech Recognition 382-384 (1993) 中所述的技術來執行。然後控制流回到步驟 506，其中該更新的叢集平均訊框被向量量化，且在步驟 510 中，以來自該測試資料庫的樣本進行收斂性測試(如同該新的潛在 SI 樣板)。本技藝的專業人士將可瞭解此遞迴處理可依需要持續到達成在該預定臨限值之上的變化及準確性結果。

一旦通過該收斂性測試(即一旦達到該臨限值)，該潛在的(現在爲最後的) SI 樣板可較佳地用於語音辨識子系

五、發明說明（ 20 ）

統，例如圖 2 的語音辨識子系統。該最後的 SI 樣板將儲存在圖 2 的樣板資料庫 108 中，或用於圖 3 的流程圖中的步驟 206。

因此，已經描述建構一演說者無關語音辨識系統的語音樣板之創新及改良的方法及裝置。本技藝的那些專業人士將可瞭解，在以上說明中所參考的資料，指令，命令，資訊，信號，位元，符號及晶片等，皆可較佳地是由電壓，電流，電磁波，磁場或粒子，光學場或粒子，或其任何組合來代表。那些專業人士將進一步瞭解配合此處所揭示的具體實施例中所描述的不同說明性邏輯方塊，模組，電路及演算法步驟，其可實施爲電子硬體，電腦軟體，或其組合。該不同的說明組件，方塊，模組，電路及步驟一般係以其功能來描述。是否該功能是實施爲硬體或軟體，係依據該特殊應用，及施加於整體系統的設計限制。專業人士可認知到在這些狀況下硬體及軟體之互換行，以及如何最佳地實施每個特殊應用所描述的功能。依照範例，配合此處所揭示的具體實施例之不同的說明邏輯方塊，模組，電路及演算法步驟，可實施或執行於一數位信號處理器(DSP)，一特定應用積體電路(ASIC)，一現場程式化閘極陣列(FPGA)或其它可程式邏輯裝置，分散閘極或電晶體邏輯，分散硬體組件，例如像是暫存器及 FIFO，一執行一組韌體指令的處理器，任何習用的可程式軟體模組及一處理器，或其任何組合來執行此處所述的功能。該處理器可較佳地是爲一微處理器，但另外，該處理器可爲任何習用

五、發明說明（ 21 ）

的處理器，控制器，微控制器，或狀態機器。該軟體模組可存在於 RAM 記憶體，快閃記憶體， ROM 記憶體，EPROM 記憶體， EEPROM 記憶體，暫存器，硬碟，可移除碟片，一 CD-ROM ，或任何其它本技藝中已知的儲存媒體形式。一範例處理器可較佳地是耦合於該儲存媒體，藉以讀取資訊或寫入資訊到該儲存媒體。另外，該儲存媒體將可整合到該處理器。該處理器及該儲存媒體可存在於一 ASIC 。該 ASIC 可存在於一電話。另外，該處理器及該儲存媒體可存在於一電話。該處理器可實施爲一 DSP 及一微處理器的組合，或成爲配合一 DSP 核心的兩個微處理器等。

因此本發明的較佳具體實施例已經顯示及說明。然而，本技藝的專業人士可以瞭解，在不背離本發明的精神或範圍之下，可對此處所揭示的具體實施例進行不同的變化。因此，本發明乃限於以下的申請專利範圍。

A5
B5

四、中文發明摘要（發明之名稱：建構一與演說者無關之語音辨識系統之語音樣板之方法及裝置）

一種建構一與演說者無關之語音辨識系統之語音樣板之方法及裝置，包含分段化一訓練發音來產生時間叢集化的段落，每個段落由一平均來代表。對於所有給定字的發音之平均值被量化來產生樣板向量。每個樣板向量與測試發音比較來產生一比較結果。該比較基本上爲一動態時間扭曲計算。該訓練發音與該樣板向量進行匹配，如果該比較結果超過至少一預定的臨限值，即產生一最佳化路徑結果，且該訓練發音根據該最佳化路徑結果來區隔化。該區隔基本上爲一K值平均分段化計算。然後該區隔的發音可以由該測試發音重新量化及重新比較，直到至少未超過一預定的臨限值。

英文發明摘要（發明之名稱："METHOD AND APPARATUS FOR CONSTRUCTING VOICE TEMPLATES FOR A SPEAKER-INDEPENDENT VOICE RECOGNITION SYSTEM"）

A method and apparatus for constructing voice templates for a speaker-independent voice recognition system includes segmenting a training utterance to generate time-clustered segments, each segment being represented by a mean. The means for all utterances of a given word are quantized to generate template vectors. Each template vector is compared with testing utterances to generate a comparison result. The comparison is typically a dynamic time warping computation. The training utterances are matched with the template vectors if the comparison result exceeds at least one predefined threshold value, to generate an optimal path result, and the training utterances are partitioned in accordance with the optimal path result. The partitioning is typically a K-means segmentation computation. The partitioned utterances may then be re-quantized and re-compared with the testing utterances until the at least one predefined threshold value is not exceeded.

（請先閱讀背面之注意事項再填寫本頁各欄）

裝 訂 線

六、申請專利範圍

1. 一種用於與演說者無關的語音辨識系統中產生語音樣板之方法，該方法包含：

分段一第一複數個發音的每個發音以爲每個發音產生複數個時間叢集的段落，每個時間叢集的段落係由一頻譜平均所代表；

對於所有的該第一複數個發音來量化該複數個頻譜平均以產生複數個樣板向量；

將該每一個樣板向量與一第二複數個發音比較來產生至少一比較結果；

當該至少一比較結果超過至少一預定的臨限值時，匹配該第一複數個發音與該複數個樣板向量，以產生一最佳的匹配路徑結果；

根據該最佳的匹配路徑結果在時間中區隔該第一複數個發音；及

重複該量化，比較，匹配及區隔，直到該至少一比較結果不會超過任何至少一個預定的臨限值。

2. 如申請專利範圍第 1 項之方法，其中該比較包含計算一變化度量。

3. 如申請專利範圍第 1 項之方法，其中該比較包含計算一準確性度量。

4. 如申請專利範圍第 1 項之方法，其中該比較包含首先計算一變化度量，其次，如果該變化度量並未超過一第一預定臨限值時，計算一準確性度量。

5. 如申請專利範圍第 4 項之方法，其中該匹配包含如果

六、申請專利範圍

該變化度量超過該第一預定臨限值或該準確性度量超過一第二預定臨限值時，匹配該第一發音與該複數個樣板向量。

6. 如申請專利範圍第 1 項之方法，其中該比較包含執行一動態時間扭曲計算。

7. 如申請專利範圍第 1 項之方法，其中該匹配包含執行一動態時間扭曲計算。

8. 如申請專利範圍第 1 項之方法，其中該匹配及該區隔包含執行一 K 值平均分段計算。

9. 如申請專利範圍第 1 項之方法，進一步包含該第一發音的終點。

10. 一種產生用於與演說者無關之語音辨識系統的語音樣板之裝置，該裝置包含：

用以分段一第一複數個發音的每個發音之裝置，以爲每個發音產生複數個時間叢集的段落，每個時間叢集的段落係由一頻譜平均所代表；

用以對於所有的該第一複數個發音來量化該複數個頻譜平均之裝置，以產生複數個樣板向量；

用以將該每一個樣板向量與一第二複數個發音比較之裝置，以產生至少一比較結果；

當該至少一比較結果超過至少一預定的臨限值時，用以匹配該第一複數個發音與該複數個樣板向量之裝置，以產生一最佳的匹配路徑結果；

用以根據該最佳的匹配路徑結果在時間中區隔該第

六、申請專利範圍

一複數個發音之裝置；及

用以重複該量化，比較，匹配及區隔之裝置，直到該至少一比較結果不會超過任何至少一個預定的臨限值。

11. 一種產生用於與演說者無關之語音辨識系統的語音樣板之裝置，該裝置包含：

分段邏輯，用來分段一第一複數個發音的每個發音，以爲每個發音產生複數個時間叢集的段落，每個時間叢集的段落係由一頻譜平均所代表；

一量化器，其耦合於該分段邏輯，並配置成對於所有的該第一複數個發音量化該複數個頻譜平均，以產生複數個樣板向量；

一收斂性測試器，其耦合於該量化器，並配置成將每一個該複數個樣板向量與一第二複數個發音比較以產生至少一比較結果；及

區隔化邏輯，其耦合於該量化器及該收斂性測試器，並配置成當該至少一比較結果超過至少一預定的臨限值時來匹配該第一複數個發音與該複數個樣板向量，用以產生一最佳匹配路徑結果，並用以根據該最佳匹配路徑結果來在時間中區隔該第一複數個發音，

其中該量化器，該收斂性測試器，及該區隔化邏輯係進一步配置成重複該量化，比較，匹配及區隔化，直到該至少一比較結果不會超過任何至少一預定的臨限值。

六、申請專利範圍

12. 如申請專利範圍第 11 項之裝置，其中該至少一比較結果爲一變化度量。

13. 如申請專利範圍第 11 項之裝置，其中該至少一比較結果爲一準確性度量。

14. 如申請專利範圍第 11 項之裝置，其中該至少一比較結果爲一變化度量及一準確性度量，其中該收斂性測試器用來首先計算該變化度量，而其次，如果該變化度量並未超過一第一預定臨限值時，計算一準確性度量。

15. 如申請專利範圍第 14 項之裝置，其中該匹配包含如果該變化度量超過該第一預定臨限值或該準確性度量超過一第二預定臨限值時，匹配該第一發音與該複數個樣板向量。

16. 如申請專利範圍第 11 項之裝置，其中該收斂性測試器用來執行一動態時間扭曲計算。

17. 如申請專利範圍第 11 項之裝置，其中該區隔邏輯用來執行一動態時間扭曲計算。

18. 如申請專利範圍第 11 項之裝置，其中該區隔邏輯包含 K 值平均語音分段計算。

19. 如申請專利範圍第 11 項之裝置，進一步包含一終點偵測器，其耦合於該分段邏輯，並用於偵測該第一發音的終點。

20. 一種產生用於與演說者無關之語音辨識系統的語音樣板之裝置，該裝置包含：

六、申請專利範圍

一處理器；及

一儲存媒體，其耦合於該處理器，並包含一組由該處理器執行的指令，以分段一第一複數個發音的每個發音來產生每個發音的複數個時間叢集段落，每個時間叢集段落由一平均值代表，量化所有的該第一複數個發音的該複數個頻譜平均來產生複數個樣板向量，比較每一個該複數個樣板向量與一第二複數個發音來產生至少一比較結果，如果該至少一比較結果超過至少一預定的臨限值時匹配該第一複數個發音與該複數個樣板向量，以產生一最佳匹配路徑結果，根據該最佳匹配路徑結果而在時間中區隔該第一複數個發音，及重複該量化，比較，匹配及區隔化，直到該至少一比較結果不超過任何至少一預定的臨限值。

21. 如申請專利範圍第20項之裝置，其中該至少一比較結果爲一變化度量。

22. 如申請專利範圍第20項之裝置，其中該至少一比較結果爲一準確性度量。

23. 如申請專利範圍第20項之裝置，其中該至少一比較結果爲一變化度量及一準確性度量，其中該組由該處理器執行的指令首先計算該變化度量，而其次，如果該變化度量並未超過一第一預定臨限值時，計算該準確性度量。

24. 如申請專利範圍第23項之裝置，其中該組指令係進一步由該處理器執行，用以如果該變化度量超過該第

六、申請專利範圍

一預定臨限值或該準確性度量超過一第二預定臨限值時，匹配該第一發音與該複數個樣板向量。

25. 如申請專利範圍第 20 項之裝置，其中該組指令由該處理器執行，用以藉由執行一動態時間扭曲計算來比較每一個該複數個樣板向量與該複數個發音。

26. 如申請專利範圍第 20 項之裝置，其中該組指令由該處理器執行來匹配區隔邏輯，其用來藉由執行一動態時間扭曲計算來匹配該第一發音與該複數個樣板向量。

27. 如申請專利範圍第 20 項之裝置，其中該組指令由該處理器執行，用以藉由執行一 K 值平均語音分段計算來區隔該第一發音。

28. 如申請專利範圍第 20 項之裝置，其中該組指令係進一步由該處理器執行來偵測該第一發音的終點。

29. 一種處理器可讀取之媒體，其包含一組處理器可執行的指令以：

分段一第一複數個發音的每個發音，以爲每個發音產生複數個時間叢集的段落，每個時間叢集的段落係由一頻譜平均所代表；

對於所有的該第一複數個發音來量化該複數個頻譜平均，以產生複數個樣板向量；

將每一個樣板向量與一第二複數個發音比較，以產生至少一比較結果；

當該至少一比較結果超過至少一預定的臨限值時，

六、申請專利範圍

匹配該第一複數個發音與該複數個樣板向量，以產生一最佳的匹配路徑結果；

根據該最佳的匹配路徑結果，在時間中區隔該第一複數個發音；及

重複該量化，比較，匹配及區隔化，直到該至少一比較結果不會超過任何至少一個預定的臨限值。

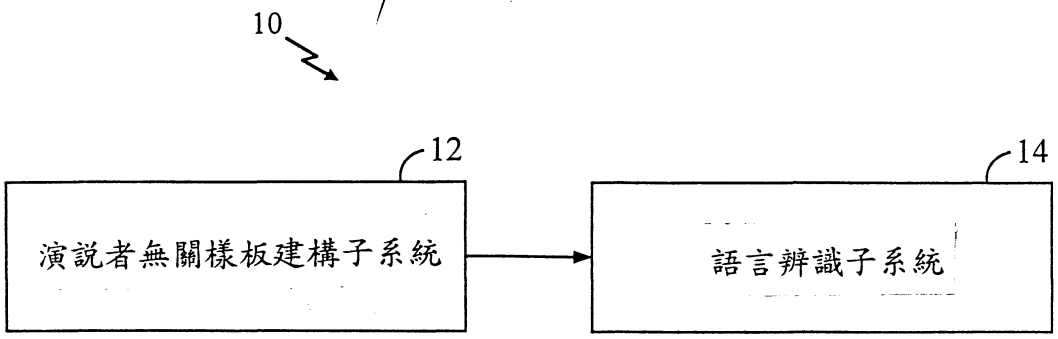
10
12
演說者無關樣板建構子系統
14
語言辨識子系統

圖 1

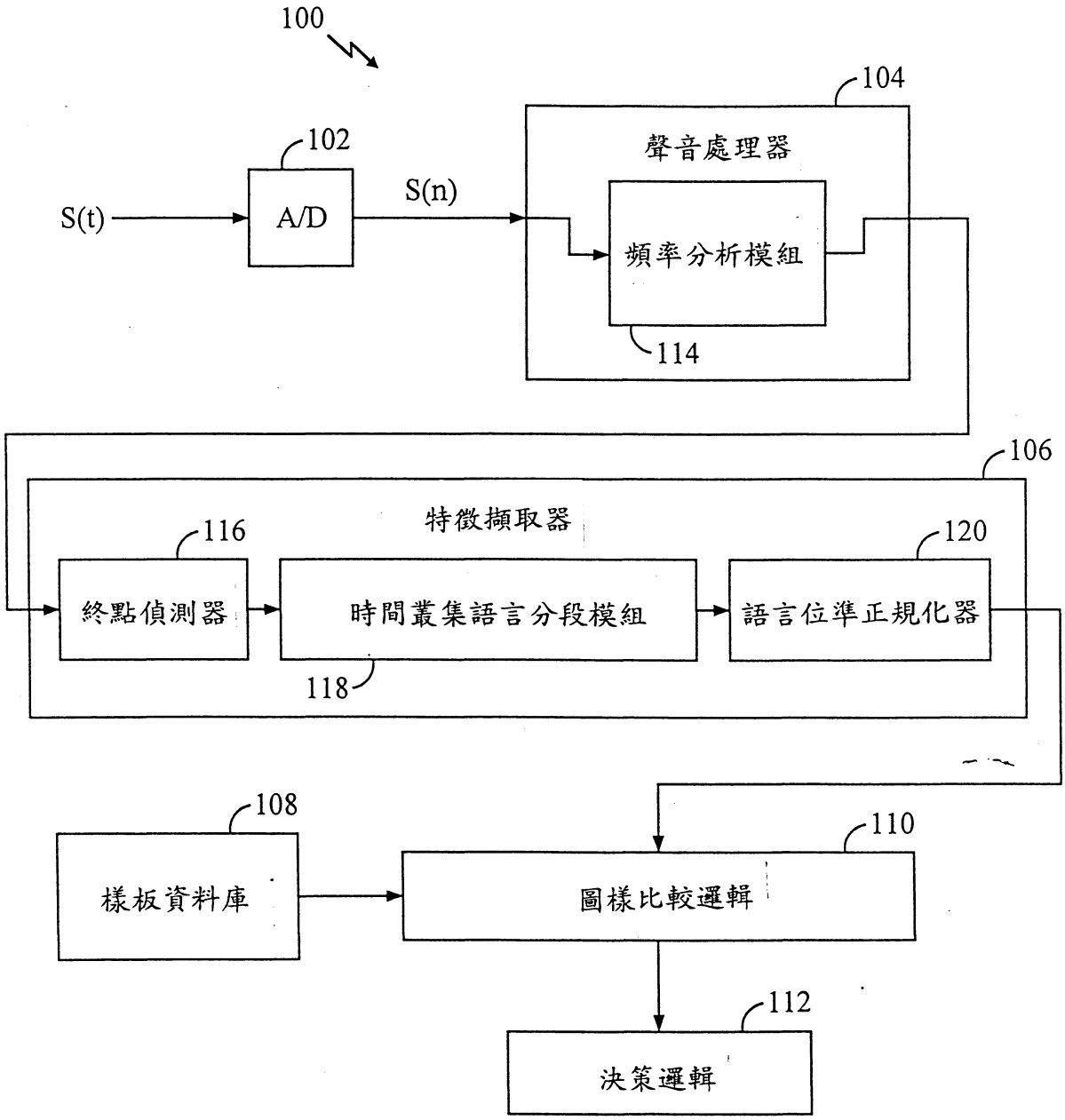
100
104
聲音處理器
102
S(t)
A/D
S(n)
頻率分析模組
114
106
特徵擷取器
116
終點偵測器
時間叢集語言分段模組
120
語言位準正規化器
118
108
樣板資料庫
110
圖樣比較邏輯
112
決策邏輯

圖 2

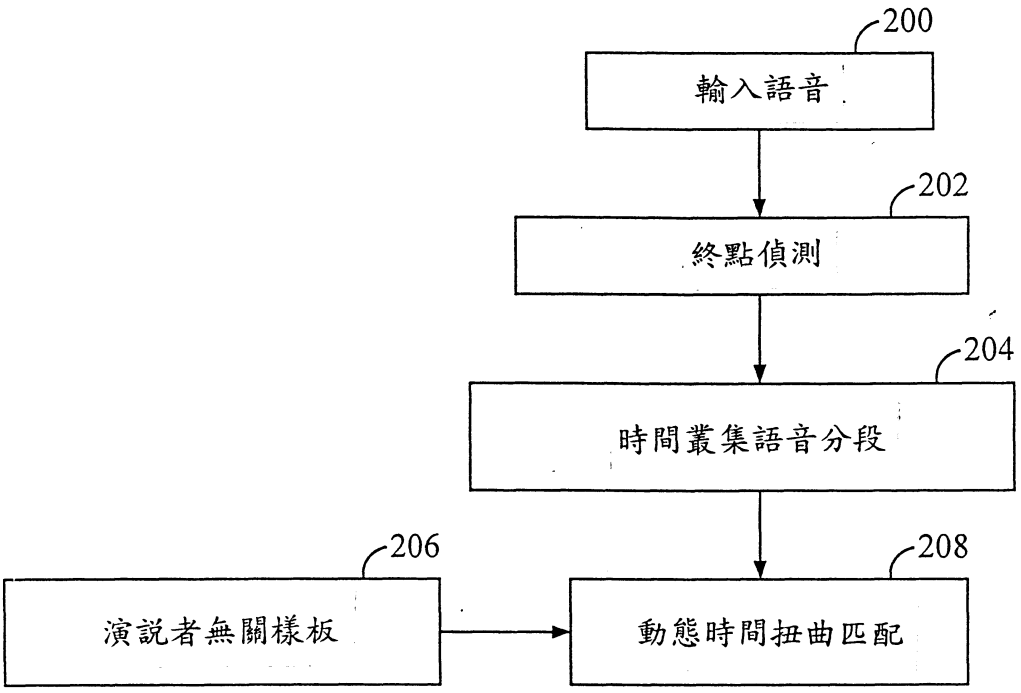
200
輸入語音
202
終點偵測
204
時間叢集語音分段
206
演說者無關樣板
208
動態時間扭曲匹配

圖 3

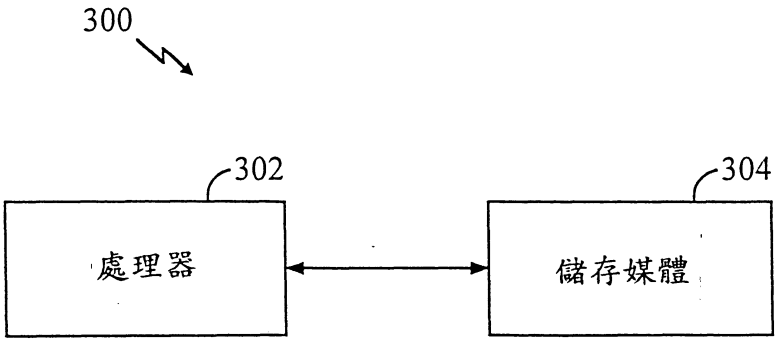
300
302
處理器
304
儲存媒體

圖 4

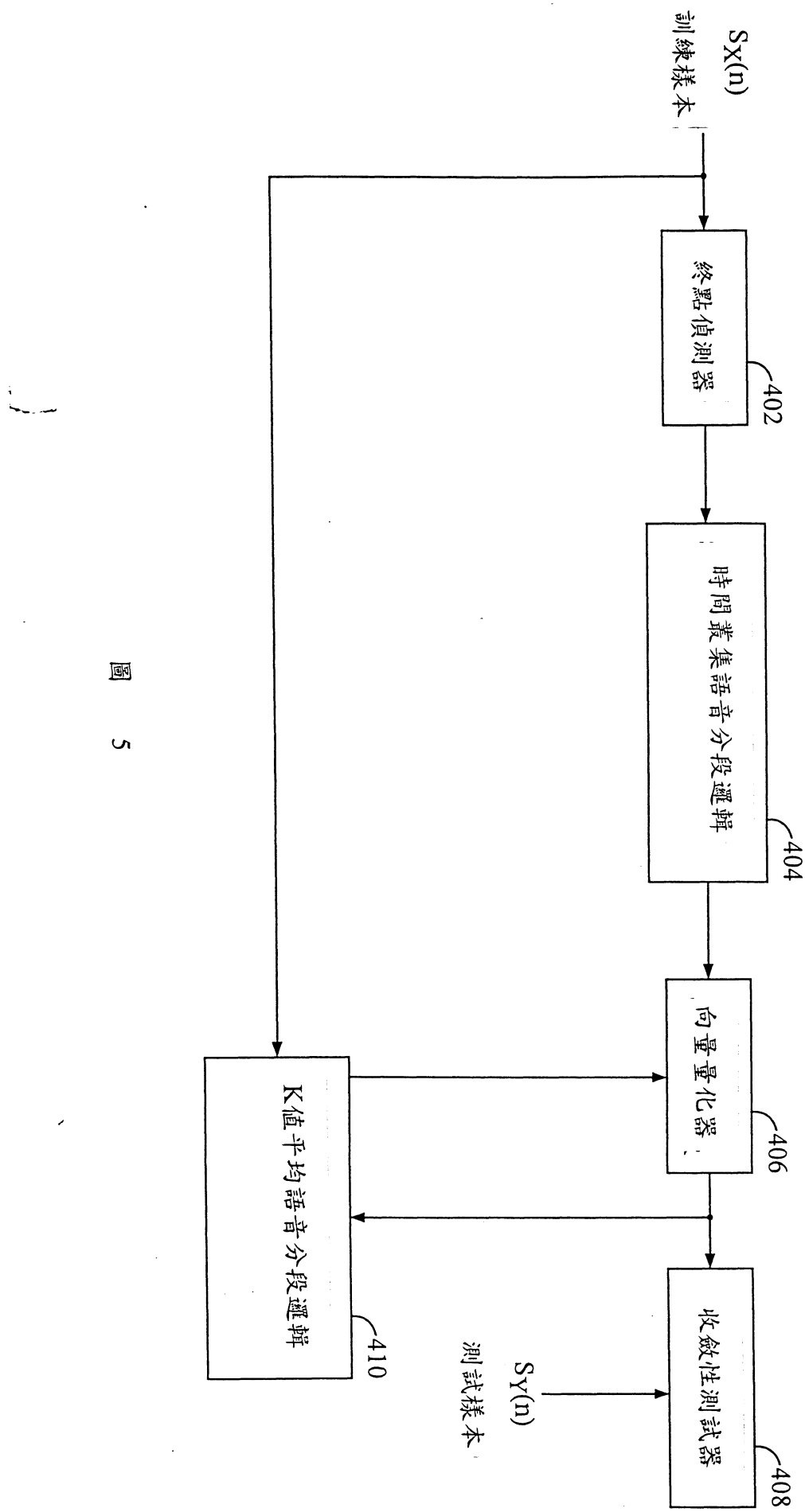
400
SX(n)
訓練樣本
終點偵測器
402
時間叢集語音分段邏輯
404
向量量化器
406
K值平均語音分段邏輯
410
收斂性測試器
408
SY(n)
測試樣本

圖 5

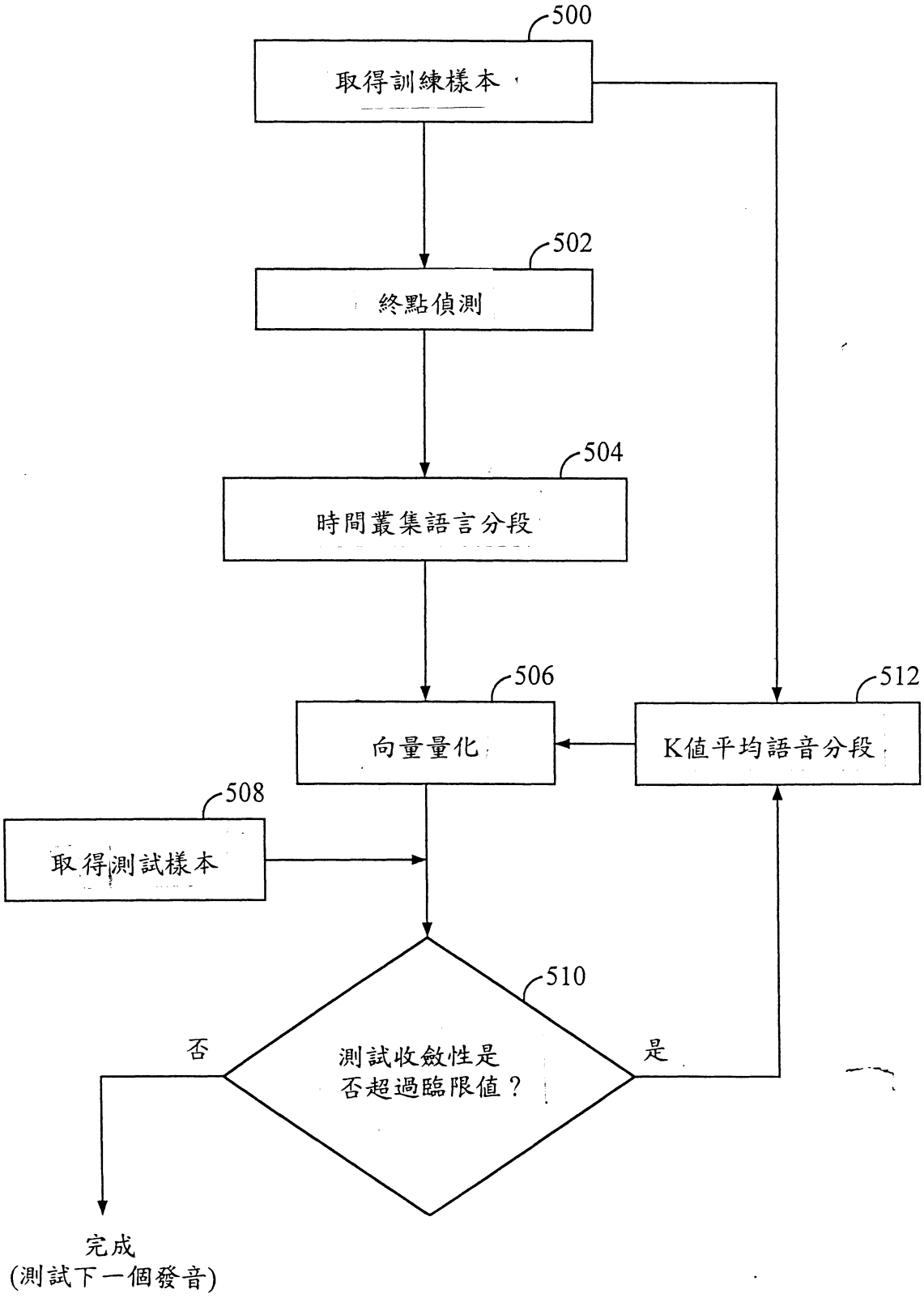
500
取得訓練樣本
502
終點偵測
504
時間叢集語言分段
506
向量量化
512
K值平均語音分段
508
取得測試樣本
510
測試收斂性是
否超過臨限值？
否
是
完成
(測試下一個發音)

圖 6